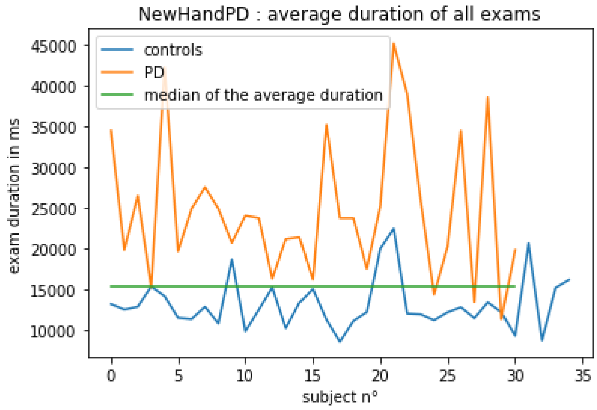
Report #6
PD-Internship
Lerner Paul
14/05/2018

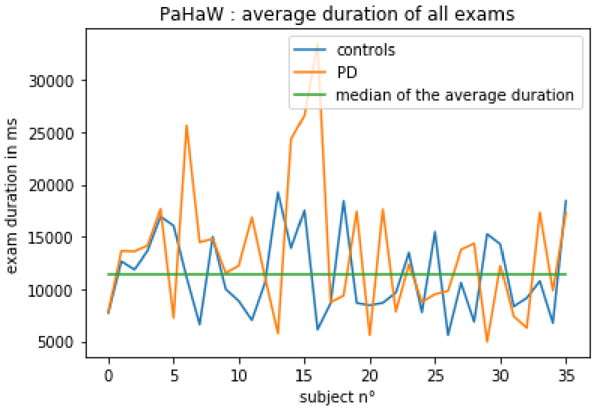
This report analyses the NewHandPD and PaHaW datasets of Pereira et al. and Drotar et al., respectively. Please refer to report #2 for analysis about related works on these datasets. Some of this report was originally present in the report on the code. I split it for better readability.

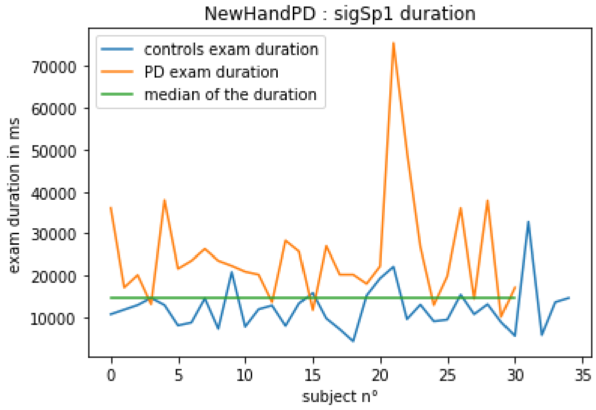
Summary

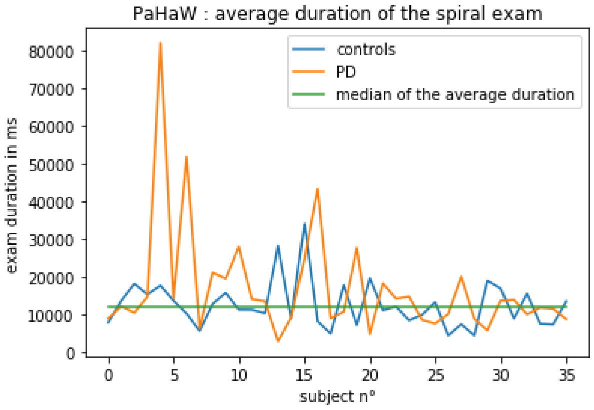
1 Data Visualization

The NewHandPD dataset has a serious defect : the PDs exams take significantly much time than Controls' exams : 11.5 seconds in average ! Thus when plotting the average duration of the exams per subject we can clearly see a threshold that separates almost all controls and all PDs. The phenomenon is visible on every single task, see below for the 1st spiral (sigSp1) exam. To have a comparison I plotted to the right the PaHaW dataset.


NewHandPD : average duration of all exams


PaHaW : average duration of all exams


NewHandPD : sigSp1 duration


PaHaW : average duration of the spiral exam

Thus when using a simple rule :

- subject is PD if exam_duration > 15242 ms *(median length over all subjects)*
- else control

I'm able to achieve 85% accuracy on the NewHandPD dataset ! See 2 Classification for proper classification tests (i.e. with train-test etc.). See 3 Interpretation for interpretations of this defect.

2 Data analysis and classification

2.1 Statistical analysis

According to the p-value, I denote test significance as null, weak, strong and very strong with _, *, ** and *** respectively.

Update on May 15th : I previously used Spearman's rank correlation coefficient measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. I used it to measure the correlation between the task duration and the targets. Although it's not actually relevant to study the correlation between a continuous and a binary variable.
So on May 15th I used Student's t-test to study the resemblance between the PD's tasks durations and the controls' tasks duration.

Before applying t-test on dataset, it needs 2 properties :

1. all input samples are from populations with equal variances.
2. data was drawn from a normal distribution.

So I applied Bartlett’s test for the first condition and Shapiro-Wilk for the second using SciPy's. As you can see from the p-values the tests are strongly significant.

Table. NewHandPD : resemblance between the PD's tasks durations and the controls' tasks duration.

task	**t-statistic**	**p-value**
circA	6.1612	5.36E-08
circB	5.9077	1.46E-07
sigDiaB	3.8808	2.49E-04
sigMea3	6.0783	7.44E-08
sigSp1	5.2307	1.99E-06
meanders	6.4848	1.48E-08
spirals	6.2149	4.33E-08
all	7.295	5.64E-10

I didn't include sigDiaA because it failed Bartlett’s test (i.e. the duration of PD's and control had different variances). Nor SigMea 1, 2, 4 nor SigSp 2, 3 and 4 because either the control or the PD group failed the Shapiro-Wilk's test (i.e. the duration was not drawn from a normal distribution.).

In order have a comparison point I studied the same thing with the PaHaW dataset. As you can see below from the p-values, none of the tests were significant, although I excluded the *l* task because it failed the Bartlett’s test. Moreover, you'll see in 2.2 Classification that it doesn't translate into good classification accuracies.

Table. PaHaW : resemblance between the PD's tasks durations and the controls' tasks duration.

task	t-statistic	t p-value
spiral	1.5826	0.118
le	1.3512	0.181
les	1.4268	0.1581
lektorka	1.421	0.1598
porovnat	0.9418	0.3495
nepopadnout	1.4159	0.1612
tram	1.0952	0.2772
all	1.5799	0.1186

2.2 Classification

To further confirm those results, I used sklearn implementation of Linear Discriminant Analysis (LDA). I feed the model only the duration of the exam (one feature).
In order to have comparable results with Afonso et al. I didn't use 10 CV but random 50-50 split with 15 runs. Pereira et al. evaluation method is unknown but it's probably similar as they have a low std (I tried 10 CV and it translated into a higher std because the test set is then only 10% of the dataset) and in a previous work on HandPD they used random 50-50 split with 10 runs.

NewHandPD : Accuracy (%) depending on the task (average over the 15 runs ± std).

task	LDA (my model)	Afonso et al.	Pereira et al. 2018
circA	80.61 ± (4.40)	76.17±6.92	68.04 ± 2.96
circB	83.43 ± (4.68)	76.69±5.38	73.41 ± 3.66
sigDiaA	75.96 ± (5.90)	68.69±7.26	73.59 ± 3.57

sigDiaB	77.17 ± (5.52)	66.30±7.38	76.32 ± 5.18
meanders	81.82 ± (3.13)	81.07±2.60	80.75 ± 2.08
spirals	82.02 ± (3.02)	81.03±2.40	78.26 ± 1.97
all	83.84 ± (3.26)	NA	95.74 ± 1.60

I averaged the lengths of the four meanders and the four spirals to provide the results of the lines meanders and spirals. Although I achieved similar results when using only one meander or one spiral to train the model. In the same way in "all" I averaged all the lengths for each subject. Majority voting provided similar results.

I displayed here the results of Afonso et al. when using first OPF then SVM because it's their best results (cf. Report #2). In the same way the results of Pereira et al. are the one where they used the "CNN-ImageNet" model on "images" of size 128x128 because it's their best results.

Notice how we outperform both of them on every single task although Pereira et al. achieved better results on all tasks (they use majority voting).

My goal here is not to outperform them but to prevent people from using this database which is obviously not challenging. Also it points out how one should be careful about age balance between patients and controls subjects.

From that I think we can conclude that the transformation of the data from sensor to image of Pereira et al. is not good or that CNN are not suited to discriminate PD and one should focus on kinematic features. Also, it does not encourage the use of Discrete Wavelet Transform as did Afonso et al. (cf. Report #2).

I used the exact same model and feature (i.e. LDA and task duration) on the PaHaW dataset. As I did before, in order to have a comparison point, I displayed here the results of Drotar et al. 2015, Dec. as it's their best results. Their result is the average over 10 CV but they don't provide for std. I advise you that, in order to get comparable results with the NewHandPD dataset, I used the same evaluation method as before (i.e. random 50-50 split with 15 runs) and not 10 CV like Drotar et al.

You can see that Drotar et al. outperform my model by a large margin on all tasks as my model barely gets above chance level on some tasks and falls behind on others.

PaHaW : Accuracy (%) depending on the task (average over the 15 runs ± std).

task	**LDA (my model)**	**Drotar et al. 2015, Dec.**
spiral	50.56 ± (6.97)	62.8

l	48.52 ± (7.23)	72.3
le	58.33 ± (6.25)	71
les	55.93 ± (3.91)	66.4
lektorka	48.70 ± (6.24)	65.2
porovnat	49.63 ± (6.16)	73.3
nepopadnout	50.19 ± (6.99)	67.6
tram	51.30 ± (5.73)	76.5
all	56.48 ± (4.61)	81.3

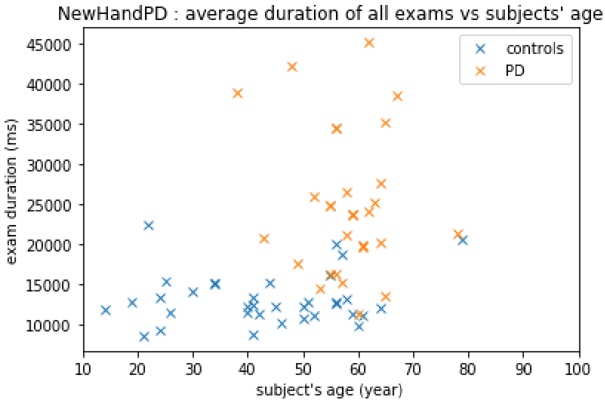
3 Interpretation

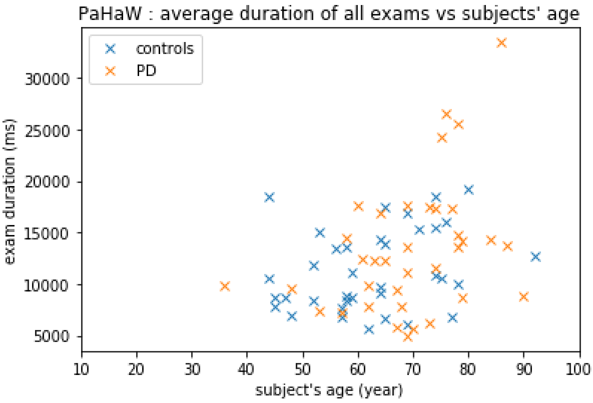
3.1 Age

In NewHandPD the controls are in average 13.8 years younger than the PDs. In report #2 we saw that age had a lot of effect on dysgraphia and that most of the datasets aimed for a equally distributed age between PDs and control. In PaHaW the controls are in average years younger than PDs. We will try to study here if these gaps are significative.

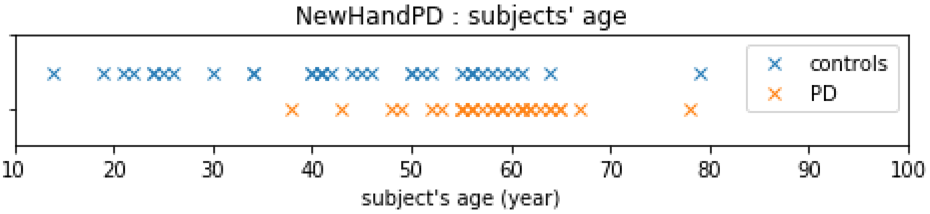
3.1.1 Visualization

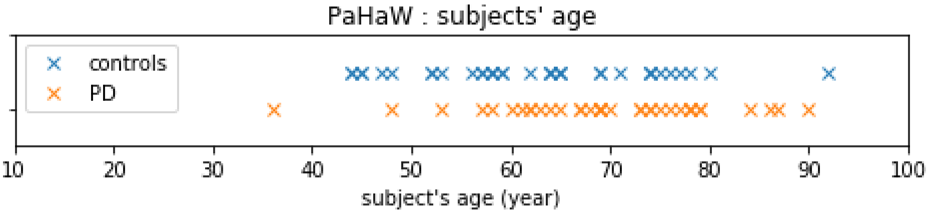
When plotting the exam duration against the subject's age, we can see two nice clusters of PDs and controls on the NewHandPD dataset, which explains why classification is so effective with only one feature : the exam duration. See the left plot below. However on the PaHaW dataset, there's only a 4 PD's cluster which explains why the same LDA model with the same feature barely gets above chance level.
You can see on the plots below you'll see in 3.1.2 Statistical analysis and classification that on both datasets, the average length of the exam is correlated to the subject's age. That's not a discovery, aging comes with this and this symptom which also come with PD. The problem is that in the NewHandPD dataset, most controls are younger than most PDs. On the contrary, you can see that on the PaHaW dataset, no control is younger than the youngest PD and no PD is older than the oldest control.





However if we focus only on the age (plotted below on two separate lines for better readability) we can still see some clusters on the PaHaW dataset, e.g. there's a group of 4 PDs who have between 80 and 90 years old (they're are not the same as the duration cluster I talked about above). This translates into a significant difference between control's and PD's age when applying a t-test and 64% accuracy when training a LDA using only the subject's age but I think it's not very significative and it's mostly due to the little number of subjects.





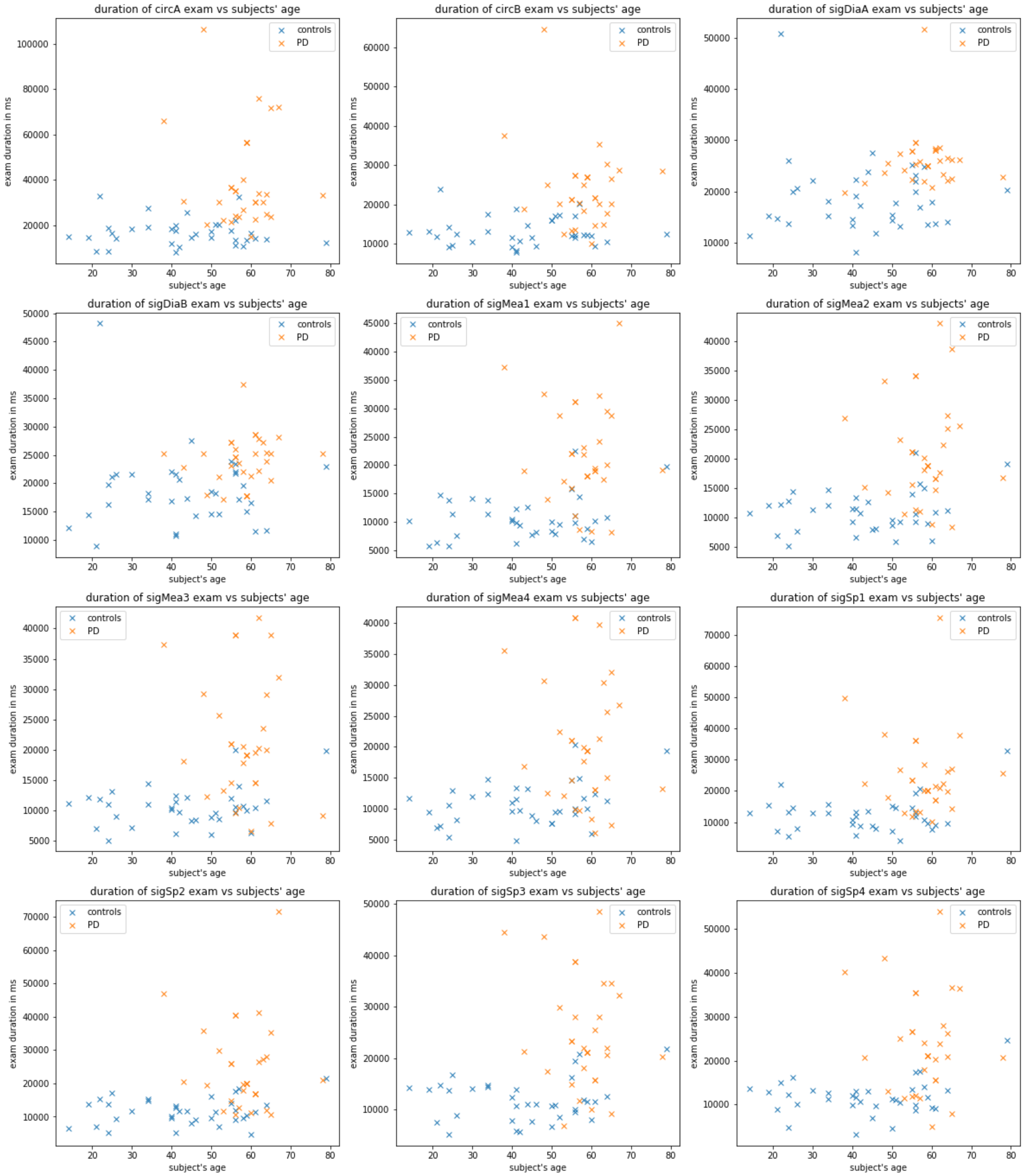


Fig. NewHandPD : duration of all exams vs. subject's age.

3.1.2 Statistical analysis and classification

I used Pearson's correlation coefficient which measures the linear correlation between two variables. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.
I used its SciPy implementation. to measure the correlation between the tasks duration and the age of the subject. As you can see below, on NewHandPD the durations of all tasks are positively correlated to the subject's age. The correlations of sigDia A and B are weaker than the other tasks. This explains why the classifier obtain worse results on these tasks (see 2.2 Classification). The correlation between task duration and subject's target was also weaker on sigDia B but not on A for some reason (see 2.1 Statistical analysis)...

Table. NewHandPD : Correlation between tasks duration and subject's age.

task	pearson correlation coefficient	p-value	significance
circA	0.2781	0.0238	**
circB	0.2533	0.0401	**
sigDiaA	0.2099	0.0907	*
sigDiaB	0.2073	0.0948	*
sigMea1	0.3822	0.0015	***
sigMea2	0.3728	0.0021	***
sigMea3	0.3265	0.0075	***
sigMea4	0.3226	0.0082	***
sigSp1	0.3079	0.0119	**
sigSp2	0.3242	0.0079	***
sigSp3	0.3169	0.0095	***
sigSp4	0.3292	0.007	***
meanders	0.3619	0.0028	***
spirals	0.3346	0.006	***
all	0.3512	0.0038	***

I performed the same with PaHaW. As we could expect from the plot and the literature, all tasks durations are significantly correlated with subject's age. See 3.1.1 Visualization for discussion.

Table. PaHaW : Correlation between tasks duration and subject's age.

task	Spearman correlation coefficient	p-value	significance
spiral	0.2163	0.0681	*
l	0.2557	0.0301	**
le	0.3088	0.0083	***
les	0.4225	0.0002	***
lektorka	0.3086	0.0084	***
porovnat	0.2787	0.0178	**
nepopadnout	0.2575	0.029	**
tram	0.2814	0.0166	**
all	0.3631	0.0017	***

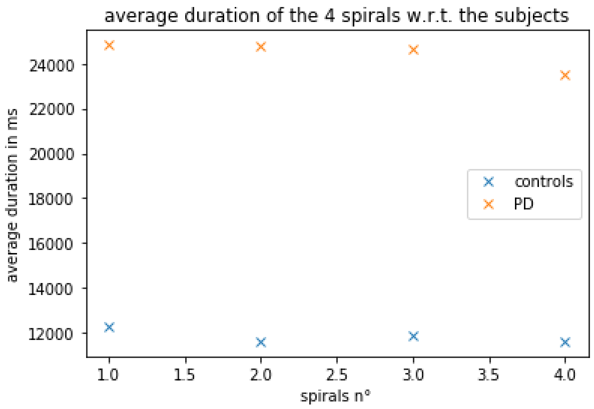
Moreover, in the same way I studied the resemblance between PDs and controls' exams' duration in 2.1 Statistical analysis, I studied here the resemblance between PDs and controls' age. Moreover, as I attempted to classify the subjects using only the duration of their exam with a LDA (see 2.2 Classification), here I used the same model using only the subject's age. See 3.1.1 Visualization for discussion. The classification evaluation is the same as before (i.e. random 50-50 split with 15 runs) therefore the result print as percentage ± std over the 15 runs.

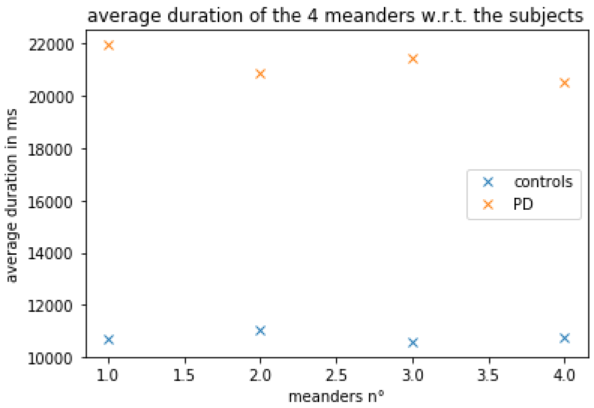
Table. Statistical Analysis and Classification of the PDs and controls' age on both datasets.

dataset	t-statistic	t p-value	significance	LDA accuracy
NewHandPD	4.8582	8.00E-06	***	75.76 ± 3.67
PaHaW	2.5606	1.26E-02	**	63.70 ± 7.49

3.2 Fatigue ?

One might ask if the duration of the exam is caused by fatigue and if PDs gets more tired than controls. Although proper statistical analysis might be required, when plotting the average duration of each 4 spirals and each 4 meanders in the order they were recorded, we cannot see any tendency (see below).





Conclusion - Todo List

I could have a look at the HandPD dataset to see if I obtain similar results.
Since there was a lot less controls in the HandPD dataset, I suspect that Pereira et al. added the youngs control we talked about here in order to balance their dataset.